Compositional Generative Modeling for Scalable Long-Horizon Robot Planning

Utkarsh A. Mishra

Fig. 1. **Long-horizon task with inter-step dependencies.** The goal of the task above is to hammer the nail into the board. A successful trajectory must handover the hammer to the right arm and both arms must coordinate to strike the nail. Compositional generative sampling can solve such tasks using only atomic skill level demonstrations and without any task-level demonstrations.

I. Motivation

Consider a dual-arm robot system driving a nail into a board as shown in Figure 1. The left arm has access to a hammer and a nail; the right arm possesses the dexterity needed for precise hammering motions. The task requires: (1) the left arm grasps the hammer, (2) hands it over to the right arm, (3) the left arm grasps the nail, and (4) the right arm strikes the nail. Suppose the robot executes steps (1) to (3) successfully by any short-horizon metric. Yet it might fail the task: if the left arm grasped the hammer by the handle and, during handover, presented it such that the right arm naturally grasped it head-first, it will deem striking geometrically infeasible. Each arm's motion was locally optimal, but their composition violated the global geometric constraint linking grasp orientation to future tool functionality.

Short-horizon optimality vs long-horizon feasibility. This scenario reveals a fundamental challenge. While learning-based methods achieve impressive short-horizon performance, they lack the ability to solve long-horizon tasks requiring inter-step dependencies. The hammer grasp orientation, irrelevant to local grasping optimality, determines whether hammering is possible three steps later. More such examples include packing multiple objects in a constrained space or rearranging objects for achieving a task goal [1].

Existing solutions to the above problem revolve around: (a) **End-to-end monolithic policies** trained on complete task demonstrations [2, 3]. However, long-horizon tasks are combinatorially many (permuting primitives yields new tasks), creating a data bottleneck. (b) **Vision-language models (VLMs)** produce semantically feasible plans ("grasp $\rightarrow$ handover $\rightarrow$ strike") but lack geometric feasibility [4]. They operate in symbolic space while robots operate in continuous geometric space. (c) **Using manually defined rules** [5] via Planning Domain Definition Language (PDDL) which are often very difficult to write for most tasks.

The compositional alternative. My research [6, 7, 8, 9, 10] leverages a key insight: while we cannot collect data for every long-horizon task, we can learn skill-level distributions from primitive demonstrations and compose them at inference to construct task-level distributions. This offers data efficiency, systematic generalization to novel skill sequences, and geometric grounding as the diffusion models operate in the continuous space where physical feasibility is determined.

II. Research Contributions

My contributions address two core questions. First, *how do we train and compose skill-level generative models* so that their joint distribution respects the geometric dependencies of long-horizon tasks? We develop frameworks that progressively tackle sequential dependencies, parallel multi-arm coordination, and trajectory stitching (Section II-A). Second, *how do we scale compositional sampling at inference time* when the composed distribution is combinatorially multi-modal? We show that embedding search within the denoising process yields adaptive inference-time scaling, and that the same principle extends to composing across planning paradigms (Section II-B) as illustrated in Figure 2.

A. Compositional Generative Modeling

The central promise of compositional planning [11, 12, 13, 14, 15] is that *we never need to collect data for and train on a complete long-horizon task*: well-learned atomic skill distributions compose to yield plans respecting geometric dependencies across the entire horizon. Diffusion models enable this through score compositionality: $\nabla \log \prod_i p_i =$

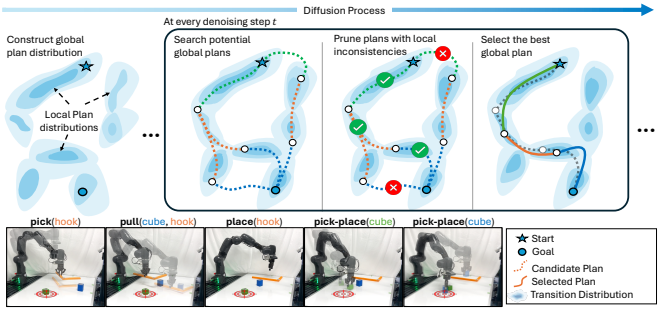

Fig. 2. **Compositional generative modeling.** Representing composition of atomic skills via the denoising diffusion process allows effective robot trajectory stitching for long-horizon planning, propagates long-horizon geometric dependencies and allows inference-time scaling with guided search.

$\sum_i \nabla \log p_i$ [16, 17] , allowing independently trained skills to compose at inference via score averaging. Given a library of skills $\pi \in \Pi$, we model each as a diffusion-based joint distribution $p_\pi(s, a, s')$ over precondition state, skill parameter, and effect state. The composition principle is straightforward: the effect of one skill overlaps with the precondition of the next, and score averaging at these overlaps propagates geometric constraints across the entire plan during diffusion denoising. This single principle, applied to progressively richer representations, addresses many core challenges.

The first challenge was whether, given a task plan, we can solve for *geometric* inter-step dependencies in continuous space. A symbolically feasible task plan specifies *what* to do but not *how*: which grasp pose enables a successful handover, which configuration enables the strike. By chaining the atomic skills through overlapping intermediate states [6], I showed that balancing forward flow (dynamics consistency) against backward flow (goal reachability) during parallel denoising resolves these dependencies automatically. The backward flow from strike skill constrains upstream grasping to favor handle-first orientations, without any complete task demonstration.

The second challenge was moving from sequential to *parallel* dependencies. Bimanual manipulations tasks require multiple manipulators acting simultaneously, coordinated through spatial constraints. By lifting the representation from sequential chaining to spatial-temporal factor graphs [7], where nodes are objects/robots and factors encode pairwise constraints, skills become temporal factors composing across both time and space. Parallel chains for each arm couple through spatial constraint factors without multi-arm training data, enabling zero-shot bimanual coordination.

Training individual models per skill does not scale to large behavior libraries. Operating instead at the *trajectory level*, subsequent work modeled trajectories as overlapping chunks with learned conditional relationships [8], enabling compositional stitching from a single unified model without any skeleton or skill labels. Finally, scaling to *visual observations* [18] by enforcing boundary agreement on denoised predictions rather than noisy states enabled training-free long-horizon video planning across unseen configurations.

B. Inference-time Scaling of Compositional Sampling

The questions above assume that score averaging produces coherent compositions. But what if we want to compose foundation-scale skill distributions representing a wide variety of short-horizon behaviors? As the diversity and multi-modality of individual skills grow, naïve score composition *mode-averages* across incompatible local modes, producing plans that are neither locally feasible nor globally coherent. Simple composition is capable but not enough to solve such complex long-horizon reasoning.

The key insight is that this can be reduced to a *search* problem [19, 20]: the correct modes exist within the skill level distribution, but finding the correct skill mode sequence requires exploring a combinatorial space. By embedding a guided search directly within the denoising process [9], maintaining a population of candidates, propagating long-range information through iterative resampling [21], and pruning incoherent candidates via likelihood-based filtering [22], compositional performance scales adaptively with inference-time compute. More compute yields better plans on harder problems including panoramic image generation and longer video generation [9].

III. Future Work

My work has progressively expanded the representation over which composition operates: from state-action transitions, to trajectories, to visual observations. In parallel, the community has converged on *World Action Models* (WAMs) [23, 24, 25] that jointly predict future video and robot actions from pre-trained video diffusion backbones. WAMs inherit spatiotemporal priors from web-scale data, enabling zero-shot generalization that eludes current VLA models [26, 27, 28]. However, they face the same horizon limitation: trained on short clips, they cannot plan over long horizons. Bridging this gap through compositional generation is the focus of my future research.

A. World Action Models as Compositional Planners

My work [9] demonstrated that short-horizon video diffusion models compose into temporally consistent long videos through score composition and guided search, suggesting that composing WAMs could yield complete executable action sequences grounded in predicted dynamics. However, generating full video frames is computationally expensive and unnecessary: the robot needs actions, not pixels. A more promising direction is to operate in the *latent space* of pretrained video models, predicting compressed visual futures and corresponding actions without decoding to full resolution. This retains the spatiotemporal priors from web-scale data while making multi-step composition tractable. My current work pursues this: learning to jointly predict video latents and actions from a pretrained video backbone, with the goal of making these predictions compositional. If each short-horizon latent-action model captures a local behavior distribution, the same score composition and guided search can chain them into long-horizon plans over compact latent representations.

B. Scaling Compositional Planning with Foundation Models

Realizing this raises open challenges. Composition in learned latent spaces must preserve the semantic and geometric structure needed for boundary agreement; unlike state space where physical meaning is explicit, latent representations require careful design to ensure score composition at boundaries produces physically meaningful constraints. Current compositional methods also assume clean skill boundaries, so composing models trained on unstructured play data requires discovering composable modes automatically from the latent-action stream. Finally, latent-space consistency does not guarantee physical feasibility or safety [10]. Solving these challenges would establish compositional generative modeling as the planning layer that transforms pretrained world models from short-horizon predictors into long-horizon planners.

References

[1] C.-F. Yang, H. Xu, T.-L. Wu, X. Gao, K.-W. Chang, and F. Gao, "Planning as in-painting: A diffusion-based embodied task planning framework for environments under uncertainty," *arXiv preprint arXiv:2312.01097*, 2023.

[2] M. Janner, Y. Du, J. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," in *International Conference on Machine Learning*, pp. 9902–9915, PMLR, 2022.

[3] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[4] K. Lin, C. Agia, T. Migimatsu, M. Pavone, and J. Bohg, "Text2motion: From natural language instructions to feasible plans," *Autonomous Robots*, vol. 47, no. 8, pp. 1345–1365, 2023.

[5] C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, and T. Lozano-Pérez, "Integrated task and motion planning," *Annual review of control, robotics, and autonomous systems*, vol. 4, no. 1, pp. 265–293, 2021.

[6] U. A. Mishra, S. Xue, Y. Chen, and D. Xu, "Generative skill chaining: Long-horizon skill planning with diffusion models," in *Conference on Robot Learning*, pp. 2905–2925, PMLR, 2023.

[7] U. A. Mishra, Y. Chen, and D. Xu, "Generative factor chaining: Coordinated manipulation with diffusion-based factor graph," in *ICRA 2024 Workshop—Back to the Future: Robot Learning Going Probabilistic*, 2024.

[8] A. Authors, "Anonymous title," *arXiv preprint*, 2025.

[9] U. A. Mishra, D. He, Y. Chen, and D. Xu, "Compositional diffusion with guided search for long-horizon planning," in *The Fourteenth International Conference on Learning Representations*, 2026.

[10] W. Jung, U. A. Mishra, N. R. Arachchige, Y. Chen, D. Xu, and S. Kousik, "Joint model-based model-free diffusion for planning with constraints," *arXiv preprint arXiv:2509.08775*, 2025.

[11] Y. Du and L. Kaelbling, "Compositional generative modeling: A single model is not all you need," *arXiv preprint arXiv:2402.01103*, 2024.

[12] D. Mahajan, M. Pezeshki, I. Mitliagkas, K. Ahuja, and P. Vincent, "Compositional risk minimization," *arXiv preprint arXiv:2410.06303*, 2024.

[13] A. Bradley, P. Nakkiran, D. Berthelot, J. Thornton, and J. M. Susskind, "Mechanisms of projective composition of diffusion models," *arXiv preprint arXiv:2502.04549*, 2025.

[14] M. Okawa, E. S. Lubana, R. Dick, and H. Tanaka, "Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[15] J. Thornton, L. Bethune, R. Zhang, A. Bradley, P. Nakkiran, and S. Zhai, "Composition and control with distilled energy diffusion models and sequential monte carlo," *arXiv preprint arXiv:2502.12786*, 2025.

[16] Y. Du, S. Li, and I. Mordatch, "Compositional visual generation with energy based models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6637–6647, 2020.

[17] Q. Zhang, J. Song, X. Huang, Y. Chen, and M.-Y. Liu, "Diffcollage: Parallel generation of large content with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10188–10198, June 2023.

[18] Y. Zhang, Y. Luo, U. A. Mishra, W. C. Shin, Y. Chen, and D. Xu, "Compositional visual planning via inference-time diffusion scaling," *arXiv preprint arXiv:2603.02646*, 2026.

[19] N. Ma, S. Tong, H. Jia, H. Hu, Y.-C. Su, M. Zhang, X. Yang, Y. Li, T. Jaakkola, X. Jia, *et al.*, "Inference-time scaling for diffusion models beyond scaling denoising steps," *arXiv preprint arXiv:2501.09732*, 2025.

[20] X. Zhang, H. Lin, H. Ye, J. Zou, J. Ma, Y. Liang, and Y. Du, "Inference-time scaling of diffusion models through classical search," *arXiv preprint arXiv:2505.23614*, 2025.

[21] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[22] A. Heng, H. Soh, *et al.*, "Out-of-distribution detection with a single unconditional diffusion model," *Advances in Neural Information Processing Systems*, vol. 37, pp. 43952–43974, 2024.

[23] J. Liang, P. Tokmakov, R. Liu, S. Sudhakar, P. Shah, R. Ambrus, and C. Vondrick, "Video generators are robot policies," *ArXiv*, vol. abs/2508.00795, 2025.

[24] M. J. Kim, Y. Gao, T.-Y. Lin, Y.-C. Lin, Y. Ge, G. Lam, P. Liang, S. Song, M.-Y. Liu, C. Finn, and J. Gu, "Cosmos policy: Fine-tuning video models for visuomotor control and planning," 2026.

[25] N. GEAR, "Dreamzero: World action models are zero-shot policies." https://dreamzero0.github.io/, 2026. Project Website.

[26] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, G. Lam, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "Openvla: An open-source vision-language-action model," *ArXiv*, vol. abs/2406.09246, 2024.

[27] M. Reuss, H. Zhou, M. Rühle, Ö. E. Yagmurlu, F. Otto, and R. Lioutikov, "Flower: Democratizing generalist robot policies with efficient vision-language-action flow policies," *ArXiv*, vol. abs/2509.04996, 2025.

[28] P. Intelligence, "π0.5: a vision-language-action model with open-world generalization," *ArXiv*, vol. abs/2504.16054, 2025.